Algorithmic Classification and Strategic Effort

JON KLEINBERG
Cornell University
and
MANISH RAGHAVAN
Cornell University

In this letter, we summarize our recent work examining the incentives produced by algorithmic decision-making. Drawing upon principal-agent models in the mechanism design literature, we construct and analyze a model of strategic behavior under algorithmic evaluation. We characterize which behaviors can be incentivized by any reasonable mechanism, showing that simple linear mechanisms are sufficient to incentivize desired behavior. However, we find that it is computationally hard to optimize even simple objectives subject to the constraint that the resulting linear mechanism induces the desired incentives.

Categories and Subject Descriptors: J.4 [**Social and Behavioral Science**]: Economics

General Terms: Economics, Theory

Additional Key Words and Phrases: Principal-agent, Strategic classification, Effort allocation

1. INTRODUCTION

This letter summarizes our recent work exploring the interaction between strategic behavior and algorithmic decision-making [Kleinberg and Raghavan 2019]. Reasoning about strategic behavior is a key aspect of game theory and mechanism design. There is a rich body of work in the economics literature on principal-agent models that considers settings in which a principal wants to incentivize an agent to exhibit certain behaviors [Grossman and Hart 1983; Holmström and Milgrom 1987; 1991; Hermalin and Katz 1991]. More recently, concerns about strategic behavior have begun to appear in the machine learning literature, especially when algorithmic decision-making systems are used to evaluate people [Dalvi et al. 2004; Brückner and Scheffer 2011; Hardt et al. 2016; Dong et al. 2018; Hu et al. 2019; Milli et al. 2019].

As we will discuss, there are key differences in the how these lines of work treat strategic behavior. As a consequence, the styles of models used differ dramatically between these fields. Our goal in this work is to develop models that integrate the disparate ideas from both of these lines of research, in order to understand how machine learning classification rules can incentivize behavior, and to characterize what kinds of behaviors can be incentivized this way. In what follows, we will formulate a model of strategic behavior under algorithmic decision-making, characterize optimal rules that decision-makers can set, and discuss the implications for data-based optimization.

Authors' addresses: kleinberg@cornell.edu, manish@cs.cornell.edu

1.1 Diverging Views of Strategic Behavior

The principal-agent and strategic machine learning literatures appear to share a common goal: how should one structure a decision-making rule to account for the strategic actions of decision subjects? However, there is a key distinction between the two. While the economics literature typically assumes that the decision-maker (typically known as the "principal") wants to incentivize particular behaviors, the computer science literature seeks to predict fixed, immutable outcomes based on strategic data, regardless of what behaviors agents exhibit.

This distinction is evident from the canonical motivating examples in the respective fields. Insurance markets are a classic example of a principal-agent setting in economics [Arrow 1963]—an automobile insurer, for example, seeks to incentivize the vehicle owner to be as careful as possible when driving, and therefore stipulates that the owner must pay a deductible on any claims. Thus, there is a particular set of behaviors (safe driving) that the insurer wants to incentivize. In contrast, strategic behavior under classification was first studied in the computer science literature in the context of spam classification [Dalvi et al. 2004]—an email either was or was not spam, and the goal was to algorithmically make this determination under the knowledge that spam authors would behave strategically to evade detection. In the machine learning context, strategic behavior is typically viewed as undesirable, and the ultimate goal is to make correct decisions in spite of agents' strategic behavior.

The distinction between economic and computer science perspectives also manifests in how the principal's (or decision-maker's) utility is modeled. In economics, agents' actions directly impact the principal's utility—if a driver gets into an accident, the insurer must pay. In computer science, there is no such relationship; agents' behavior impacts the decision-maker only insofar as it makes it difficult for them to make correct decisions. An email provider doesn't intrinsically care about which phrases or punctuation marks a spam author inserts to bypass spam filters, as long as they are able to correctly identify the message as spam.

1.2 Making Decisions about People

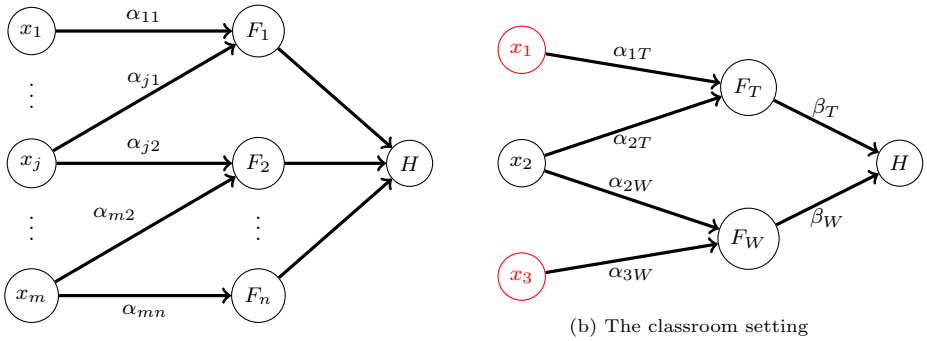
In the case of spam classification, it would seem obvious that we should think of strategic behavior as simply making the task harder without affecting the ground truth of whether or not an email is spam. But algorithms are increasingly used in contexts where things are not so clear. In credit scoring, for example, an individual's credit history is reduced to a single number by a fixed decision rule—an algorithm, of sorts [Citron and Pasquale 2014]. Because credit scores are powerful determinants of an individual's ability to participate in the financial system, people engage in a variety of "credit repair" exercises in order to improve these scores. This is clearly strategic behavior, and while some credit repair techniques may not change underlying creditworthiness, some strategic behavior (e.g., proactively paying down debt) may actually increase an individual's ability to pay back a loan.

Thus, when making decisions or predictions about people, we cannot simply treat the ground truth as fixed. In some cases, individuals' strategic behavior can change the outcomes that algorithmic decision-making systems seek to predict. In order to capture these cases, we propose a model that applies ideas from principal-agent theory to the algorithmic decision-making setting.

2. BUILDING A MODEL

2.1 Effort, Features, and Evaluation

A key feature of models of strategic behavior in both the principal-agent and strategic classification literatures is that of hidden actions: decision-makers can't observe the exact actions taken by a decision-subject. In the principal-agent literature, this is typically captured through stochastic outcomes, where each action by the agent produces a distribution over possible outcomes, but only one outcome is realized and observed by the principal. In the strategic classification literature, actions are typically not explicitly modeled; instead, a decision-maker simply observes a feature vector, where the values of the feature vector can be strategically manipulated (to a limited degree) by the decision subject. Our model will require decisions to be made based on observable feature vectors, while allowing decision-makers to still have preferences over the underlying behaviors taken by decision subjects. Given multiple possible ways of manipulating feature vectors, our model will allow decision-makers to structure their decision rules to incentivize certain behaviors over others.



(a) General model

(b) The classroom setting

Fig. 1: The conversion of effort to feature values can be represented using a weighted bipartite graph, where effort x_j spent on action j has an edge of weight α_{ji} to feature F_i.

We model the relationship between actions and observations as what we call an *effort graph*, shown in Figure 1a. The effort graph is a bipartite graph with two sets of vertices:

(1) effort variables $x_1, \ldots, x_m$, which specify the m actions an agent can take, and
(2) features $F_1, \ldots, F_n$, which specify the n values that the decision-maker actually observes.

The edges between effort variables and features specify the relationships between actions and observations. In particular, each edge between x_j and F_i contains a parameter α_{ji} indicating the degree to which action j increases feature i. We write this formally as

$$F_i = f_i\left(\sum_{j=1}^{m} \alpha_{ji} x_j\right), \tag{1}$$

where f_i is some continuous, concave function. Thus, the larger α_{ji} is, the more action x_j increases the value of feature F_i. We assume each $\alpha_{ji} \geq 0$, meaning effort can only increase the value of a feature. We also assume that both the decision-maker and decision subjects have full knowledge of the model's structure and parameters.

The decision-maker must evaluate the decision subject by giving them a score $H = M(\vec{F})$. We assume that the features are "positively oriented," meaning higher feature values are better, from the decision-maker's perspective. Thus, we will restrict the mechanism M to be monotone, meaning higher feature values can never result in worse outcomes for the decision subject. We assume that decision subjects know how they will be evaluated, i.e., they know M. Their goal is to maximize the score they receive by allocating their effort over the effort variables $x_1, \ldots, x_m$. Given a budget B, the agent will allocate their effort across the m effort variables $x_1, \ldots, x_m$ as[1]

$$\max_{x_1,\ldots,x_m} M(\vec{F}) \qquad \text{s.t. } F_i = f_i\left(\sum_{j=1}^{m} \alpha_{ji} x_j\right) \quad (\forall i)$$

$$\sum_{j=1}^{m} x_j \leq B$$

$$x_j \geq 0 \quad (\forall j)$$

For an arbitrary mechanism M, this optimization problem is intractable; however, note that in the special case where M is linear (i.e., a weighted sum of the F_i's), this is a convex optimization.

Thus, our model incorporates strategic effort that can produce beneficial outcomes as in models from the principal-agent literature. In contrast with these earlier lines of work, however, our model also captures the additional notion that multiple activities can be simultaneously incentivized, leading to different means of producing the same feature vector in ways that will be indistinguishable to the principal. This indistinguishability is key to a number of the motivating classification examples.

With this model, we can reason about how an agent will respond to a particular mechanism M. This allows us to ask a number of basic questions about the feasibility of incentivizing desired behaviors. When can a particular behavior or set of behaviors be incentivized? What sorts of mechanisms do so? What is the structure of the space of incentivizable behaviors? We summarize our main results on these questions in Section 3. First, we provide a concrete instantiation of the model to make these questions more concrete.

2.2 Instantiating the Model

Here, we present a simple example of how this model can be instantiated in the context of a classroom setting. Suppose that there are two graded assignments in

[1]While we present the agent as having a fixed budget, our results are largely unchanged in the setting where the agent has a fixed cost per unit effort.

a class: a test F_T and a homework assignment F_W. Students have three possible actions: cheat on the test (x_1), study the material (x_2), and cheat on the homework (x_3). Cheating on the test contributes to a student's test score, cheating on the homework contributes to their homework score, and studying the material contributes to both scores. Figure 1b depicts this setting with the associated parameters α_{ji}, which govern the degrees to which each action improves each feature. The teacher has to assign a final grade based on the assignment scores F_T and F_W. For simplicity, assume that the final grade H must be a linear combination of the two: $H = \beta_T F_T + \beta_W F_W$, where $\beta_T, \beta_W \geq 0$. How can the teacher incentivize students to study instead of cheat?[2]

Intuitively, whether or not this is possible depends on the relative difficulty of studying as opposed to cheating: when cheating is much more efficient than studying, it is impossible to incentivize the student to study purely by manipulating the grading weights β_T and β_W. Conversely, when studying is more efficient than each form of cheating, the student will always be incentivized to study. Somewhat surprisingly, however, if studying is only slightly less efficient than cheating, it may still be possible to incentivize studying. For example, when $\alpha_{2T} = \alpha_{2W} = 2$, $\alpha_{1T} = \alpha_{3W} = 3$, and the effort conversion functions are the same ($f_T = f_W$), then setting $\beta_T = \beta_W = 1$ will incentivize the agent to invest their entire budget into studying (x_2), as depicted in Figure 2.

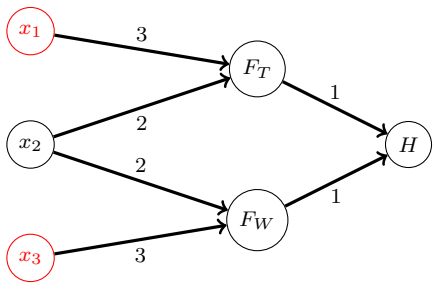


Fig. 2: An instantiation of the classroom setting, where x_2 (studying) is incentivized even though it is less efficient than each individual form of cheating (x_1 and x_3).

More generally, our main results imply that when $\frac{\alpha_{2T}}{\alpha_{1T}} + \frac{\alpha_{2W}}{\alpha_{3W}} \geq 1$, then there must exist some β_T, β_W that incentivize studying over cheating. Of course, the teacher might decide to use a complex non-linear grading scheme; however, our results imply that whenever *any* monotone mechanism incentivizes studying over cheating, there exists a linear mechanism that does so as well. In the following section, we formally state our results.

[2]We have left some parameters (the budget B and the effort conversion functions f_T and f_W) uninstantiated, as the main conclusions of our example will not depend upon them.

3. RESULTS

3.1 Incentivizing Desirable Behavior

With this intuition, we are ready to summarize our main results. Our first theorem characterizes when an agent can be induced to allocate their budget according to a particular target "effort profile," a vector in $\mathbb{R}^m_{\geq 0}$ with entries that sum to B. Let $\mathcal{S}(x)$ be the support (set of nonzero entries) of x.

THEOREM 3.1. *For an effort graph G and an effort profile $x^* \in \mathbb{R}^m_{\geq 0}$, the following are equivalent:*

(1) *There exists a linear mechanism that incentivizes x^*.*
(2) *There exists a monotone mechanism that incentivizes x^*.*
(3) *For all x such that $\mathcal{S}(x) \subseteq \mathcal{S}(x^*)$, there exists a linear mechanism that incentivizes x.*

Furthermore, there is a polynomial time algorithm that decides the incentivizability of x^* and provides a linear mechanism β to incentivize x^* whenever such β exists.

Theorem 3.1 implies that in our setting linear mechanisms are optimal in the following sense: whenever a "reasonable" mechanism can incentivize a particular behavior, there is a linear mechanism that can do so.

3.2 Optimizing Objectives

Typically, algorithmic decision rules don't exist solely for the purpose of incentivizing particular behaviors. They often seek to optimize predictions of certain outcomes, perhaps based on some historical data. However, our second theorem shows that it may be hard to do so while maintaining the desired incentives. Let g be a concave function over the space of effort profiles, and let D be the set of "desirable" actions in which the evaluator is willing to incentivize the agent to invest effort. Let $\mathcal{X}_D = \{x \mid \mathcal{S}(x) \subseteq D\}$, i.e., $\mathcal{X}_D$ is the set of effort profiles the evaluator is willing to incentivize.

THEOREM 3.2. *Consider the optimization problem*

$$\max_{x \in \mathcal{X}_D} g(x) \text{ s.t. } x \text{ is incentivizable.} \tag{2}$$

(1) *If there exists an x^* such that $\mathcal{S}(x^*) = D$ and x^* is incentivizable, then (2) can be optimized in polynomial time.*
(2) *If $|D|$ is constant, then (2) can be optimized in polynomial time.*
(3) *In general, (2) is NP-hard to optimize.*

Thus, Theorem 3.2 tells us that it can be hard to incentivize a complex set of behaviors while also optimizing other objectives.

4. DISCUSSION

In this letter, we have presented our analysis of a model of strategic behavior under decision-making, combining perspectives from economics and computer science. While our model makes several assumptions, a number of concurrent and subsequent papers have sought to relax some of these assumptions and further explore

the interface between principal-agent and machine learning settings. For example, while our model focuses on a particular agent, Alon et al. and Haghtalab et al. consider multi-agent settings [Alon et al. 2020; Haghtalab et al. 2020]. We also assume that the relationships between actions and features are known; Miller et al., Bevhavod et al., and Shavit et al. discuss the problem of inferring these causal relationships [Bechavod et al. 2020; Shavit et al. 2020; Miller et al. 2019]. Tang et al. consider a partial information setting, characterizing solutions that are robust to incomplete knowledge of action-feature relationships [Tang et al. 2020]. Dütting et al. study the gap between simple and optimal contracts in a more general principal-agent setting [Dütting et al. 2019]. Taken together, this emerging body of research highlights the importance of incentives in algorithmic decision-making: by accounting for incentives, we can increase the welfare of decision-makers and decision subjects alike.

REFERENCES

Alon, T., Dobson, M., Procaccia, A. D., Talgam-Cohen, I., and Tucker-Foltz, J. 2020. Multiagent evaluation mechanisms. In *AAAI*. 1774–1781.

Arrow, K. J. 1963. Uncertainty and the welfare economics of medical care. *American Economic Review 53,* 5, 941–73.

Bechavod, Y., Ligett, K., Wu, Z. S., and Ziani, J. 2020. Causal feature discovery through strategic modification. *arXiv preprint arXiv:2002.07024*.

Brückner, M. and Scheffer, T. 2011. Stackelberg games for adversarial prediction problems. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 547–555.

Citron, D. K. and Pasquale, F. 2014. The scored society: Due process for automated predictions. *Wash. L. Rev. 89*, 1.

Dalvi, N. N., Domingos, P. M., Mausam, Sanghai, S. K., and Verma, D. 2004. Adversarial classification. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 99–108.

Dong, J., Roth, A., Schutzman, Z., Waggoner, B., and Wu, Z. S. 2018. Strategic classification from revealed preferences. In *Proceedings of the 2018 ACM Conference on Economics and Computation*. 55–70.

Dütting, P., Roughgarden, T., and Talgam-Cohen, I. 2019. Simple versus optimal contracts. In *Proceedings of the 2019 ACM Conference on Economics and Computation*. 369–387.

Grossman, S. J. and Hart, O. D. 1983. An analysis of the principal-agent problem. *Econometrica 51,* 1, 7–45.

Haghtalab, N., Immorlica, N., Lucier, B., and Wang, J. 2020. Maximizing welfare with incentive-aware evaluation mechanisms. In *29th International Joint Conference on Artificial Intelligence*.

Hardt, M., Megiddo, N., Papadimitriou, C. H., and Wootters, M. 2016. Strategic classification. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*. 111–122.

Hermalin, B. E. and Katz, M. L. 1991. Moral hazard and verifiability: The effects of renegotiation in agency. *Econometrica 59,* 6, 1735–1753.

Holmström, B. and Milgrom, P. 1987. Aggregation and linearity in the provision of intertemporal incentives. *Econometrica 55,* 2, 303–328.

Holmström, B. and Milgrom, P. 1991. Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design. *JL Econ. & Org. 7*, 24.

Hu, L., Immorlica, N., and Vaughan, J. W. 2019. The disparate effects of strategic manipulation. In *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019*. 259–268.

Kleinberg, J. and Raghavan, M. 2019. How do classifiers induce agents to invest effort strategically? In *Proceedings of the 2019 ACM Conference on Economics and Computation*. 825–844.

Miller, J., Milli, S., and Hardt, M. 2019. Strategic classification is causal modeling in disguise. *arXiv*, arXiv–1910.

Milli, S., Miller, J., Dragan, A. D., and Hardt, M. 2019. The social cost of strategic classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019*. 230–239.

Shavit, Y., Edelman, B., and Axelrod, B. 2020. Causal strategic linear regression. In *Proceedings of the 37th International Conference on Machine Learning*.

Tang, W., Ho, C.-J., and Liu, Y. 2020. Linear models are robust optimal under strategic behavior.